The Evaluation Plan for the ACE 2007 Pilot Evaluation of Entity Translation

1 INTRODUCTION

A pilot evaluation of "Entity Translation" will be conducted in early 2007. It will be coordinated along with the ACE 2007 information extraction evaluation, although the registration and evaluation processes will be completely separate. Participants in the Entity Translation (ET) pilot evaluation have no obligation to participate in the separate monolingual ACE information extraction evaluation.

2 TASK DEFINITION AND MOTIVATION

Systems participating in the pilot Entity Translation (ET) task will be evaluated on their ability to take in a text document in one language (either Mandarin Chinese or Arabic) and emit an English language catalog of the entities mentioned in the document, as well as a normalized version of any date/time expressions. Performance will be measured based on a number of parameters, including coverage of the entities recognized as well as the quality of the English language renderings of each entity's mentions. The types of mentions that are to be included in this reporting include those that are in the form of names, descriptors (nominal phrases) and pronouns.

The organizers of this evaluation wish to focus technical efforts on the challenge of capturing the essential information content of the foreign language documents in English, and less on where and how this information was presented in the original foreign language document. For this reason the scoring metric will not rely on systems identifying where in the source language document the English language information was found.

This evaluation can be viewed as one point in a spectrum of "cross-language information extraction" tasks. While the current pilot evaluation will focus on capturing information about the entities and temporal expressions, it is expected that future evaluations will expand to include other types of information (as described in the existing monolingual ACE tasks -- values, relations and events).

Much of the evaluation infrastructure in this evaluation (data formats, general approach to scoring, etc.) will borrow from the ACE Entity Detection and Recognition (EDR) evaluation. There are a number of reasons for this. The data captured in the ACE EDR evaluations have been reviewed and tested by organizers and system developers over the course of a number of years, so the guidelines are highly refined for all of the languages in which this task has been annotated. The information extracted in the ACE EDR data has been determined to be of practical value for a range of applications, and by using it within the context of this cross-language evaluation the community can better evaluate its impact. Finally, by adopting these same data formats and associated scoring mechanisms we are able to reduce the cost and complexity of initiating this new evaluation in this first, "pilot" setting.

2.1 EXTRACTING AND TRANSLATING ENTITIES AND TEMPORAL EXPRESSION NORMALIZATION

The set of entity types to be recognized, as well as the information expected to be captured about each entity, is identical to the information required for the standard (monolingual) ACE Entity Detection and Recognition (EDR) task. Participants should consult the ACE 2007 Evaluation Plan (the NIST ACE web site is at http://www.nist.gov/speech/tests/ace/ace07/) for the details on what information EDR systems are expected to capture about a source document. There are three main differences between mono-lingual EDR and the cross-language Entity Translation (ET) task:

(1) ET requires that foreign language temporal expressions be recognized and translated into English, as well as have the "normalized" values of these temporal expressions captured in the Timex2 format, as defined in the ACE Temporal Expression Recognition and Normalization task (TERN);

(2) Systems are not required to specify where in the source document entity and temporal expression mentions are found (see Sec. 3 and the ACE 2007 Evaluation Plan for APF formatting specifics); and

(3) While systems are required to produce a single best translation and/or transliteration for each name mention and attribute, the reference data against which systems are evaluated will include name variants to accommodate the fact that there is recognized diversity in how names in foreign scripts can be legitimately rendered in English.

To learn about the details of the ACE EDR and TERN task, including the data formats and scoring metrics, consult the ACE 2007 evaluation web site cited above.

Unlike for the mono-lingual ACE EDR task, systems performing the Entity Translation (ET) task are not required to identify where in the Chinese or Arabic source document a given entity mention or temporal expression has been detected. There is a philosophical motivation for this, as well as a practical one. The philosophical motivation is that the organizers wish to gradually move the information extraction task into a "database filling" or "knowledge based" model, where evaluation will consist of measuring the state of a database (knowledge base) after some amount of source text has been processed. Eventually this could include cross-document entity disambiguation and normalization, as well as the co-reference of relations and events that are cross referenced in multiple texts.

A practical motivation for having systems be able to avoid specifying mention locations in the source texts is that the organizers anticipate that some system developers may choose to approach this task as a processing sequence in which automatic machine translation of the complete source text is followed by more-or-less standard ACE EDR processing of the English language text. In this approach the systems would not easily be able to identify where entity mentions actually occur in the original foreign language source document.

The system output format requirements for Entity Translation retain the identical XML structures as required for monolingual EDR, (but see the section below on data formats (Sec. 3) for the details).

2.2 DIAGNOSTIC TESTS

In order to encourage the participation of organization who may have developed approaches to only parts of the overall Entity

Translation task, the evaluation will also incorporate a "diagnostic" evaluation track, in which reference entity extraction information on the original source document (either Chinese or Arabic) will be made available to the participants. Systems will then be evaluated on the Entity Translation task in the identical manner.

3 DATA FORMATS

The data format required for system output is identical to that required for the 2007 ACE EDR evaluation. For each source document in the evaluation data set the system is required to produce a single output file in the APF xml format. However, unlike in the monolingual ACE EDR task, the values of the CHARSEQ element attributes START and END are completely ignored for the ET task. Systems are expected to have these CHARSEQ element attributes present in their system output, but their values may be any integer.

4 EVALUATION

The scoring procedure for the Entity Translation task is derived directly from its counterpart in the monolingual ACE EDR evaluation task. It shares the same evaluation parameter matrix, and it has a very similar value optimizing entity mapping mechanism. However, the ET scoring procedure evaluates the quality of individual mentions differently than in EDR, due to the absence of offset information, and it includes a new evaluation of name attributes.

Mention Scoring. The ET scorer will evaluate the quality of a system generated mention by using one of two metrics. Entity mentions of type NOM or PRO will use the Meteor automatic machine translation evaluation tool to compute the degree to which a given system mention matches a reference mention. The score for system entity mentions of type NAM will be computed by a more complex method that treats the name mention head differently from that portion of the mention outside the scope of the mention head. The **head** will be scored using the Meteor scorer by comparing the system head with each distinct name variant of the reference entity (as captured in the name entity_attributes elements in APF). The head score will be the maximum Meteor score computed for these variants. The **extent** will be scored using the Meteor scorer to compare the system extent with the reference extent, as before. Before comparison, however, both the ref and the sys extents will be modified by eliminating (the first occurrence of) the ref (or sys) head character string in the ref (or sys) extent, respectively, using the Perl substitution operator ($extent =~ s/$head//). The NAM mention score will be the average of the head score and extent score.

Name Attribute Scoring. The ACE APF system output format (as used for both EDR and ET) captures distinct name mention head strings as "entity attribute" information. For ET the quality of these name translations/transliterations will be evaluated separately from the name mentions. For each putative mapping between a reference entity and a system output entity, the entity score will be multiplied by a name matching factor (NMF) that reflects the number of system entity names that do and do not match the reference entity names (including one or more name variants for each distinct name). Note that in order for names to match they must match exactly, character by character (though case will be disregarded). NMF is the product of a factor representing the number of distinct system names that have matching reference names (NMF_match), and a factor representing the number of distinct system names that don't have matching reference names (NMF_wrong). The value of these two contributing factors is a function of the number of system name attributes that match or do not match, respectively, as shown in this table:

n	NMF_match	NMF_wrong
0	0.50	1.00
1	1.00	0.75
2	1.10	0.60
3	1.15	0.55
4 or more	1.20	0.50

The ET reference data are derived from a single expert translation of the original (Chinese or Arabic) source document, enriched with alternative acceptable name variants. The name variants will be developed by expert linguists examining the names generated by system submissions, and possibly other resources. The linguists will be instructed to accept English name variants that would be completely acceptable to an English-only consumer of these data. No single "standard" or convention will be imposed. A key requirement placed on acceptable name translations are that they "preserve meaning" in English. That is, they should allow a reasonably informed monolingual reader to identify the referent. Linguists will be instructed not to accept "near misses," but only those name variants that they would deem appropriate for a "well edited" report. The linguists responsible for the compilation of these name variants will work against a set of guidelines that will be developed and refined in the course of examining data received from system submissions and elsewhere. These guidelines, entitled "2007 Entity Translation Name Variant Guidelines," will be published and updated intermittently on the NIST web site.

4.1 THE 2007 EVALUATION CORPUS

The source documents employed in both tracks (Chinese, Arabic) of the Entity Translation evaluation have been obtained from two distinct types of information sources: some of the data come from newswire sources, and some of the data come from weblogs ("blogs") of one form or another. One third of the foreign language source documents are expert translations from documents written in the other language, the second third comes from expert translations of documents originally written in English, and the final third were authored in the source language.

Table 1 The ET07 evaluation corpus statistics.

Source	Test epoch	Approximate size
Native Arabic Resources		
Newswire	1/01	7,250 words
Weblog	3/05 – 4/05	7,900 words
Native Chinese Resources *(1.5 characters = 1 word)*		
Newswire	1/01	11,600 words
Weblog	3/05 – 4/05	3,400 words
Native English Resources		
Newswire	Jul-Aug/03	10,400 words
Weblog	Mar-Apr/05	7,300 words

A participant in the Arabic track would therefore expect approximately 7,250 + 11,600 + 10,400 = 29,250 words of Arabic newswire data.

4.2 ENTITY TRANSLATION EVALUATION SOFTWARE

System output can be scored by individual sites by downloading and using the ACE ET scorer. This scoring script is written in perl and takes two files as input: the reference data (in APF format) and the system file (also in APF format). The latest version of this scoring script can be downloaded from the NIST ACE web site:

http://www.nist.gov/speech/tests/ace/ace07/software.htm

4.3 EVALUATION SCHEDULE

Below is a schedule for the ACE evaluation,
with important dates associated with the
Entity Translation (ET) task highlighted in **red**.

Date	Event
Jan. 19, 2007	Deadline to register[1] for participation in the ACE07 and ACE-ET07 evaluations
Before Jan. 26, 2007	Organizations participating in ET must submit sample system output to test format conformance
Jan. 29, 2007	ACE07 Arabic evaluation day ACE ET Chinese and Arabic data released
Jan. 30, 2007	ACE07 Chinese evaluation day
Jan. 31, 2007	ACE07 English evaluation day
Feb. 01, 2007	ACE07 Spanish evaluation day
Jan. 29-Feb. 09, 2007	*Entity Translation evaluation period*
Feb. 12, 2007	Ground-truth entity mentions available for diagnostic EDR task
Feb. 14, 2007	(noon deadline, EST) Diagnostic EDR results due at NIST
Feb. 14, 2007	Ground-truth ENTITIES available for diagnostic RDR , VDR and ET tasks
Feb. 16, 2007	(noon deadline, EST) Diagnostic RDR, VDR and ET results due at NIST
Feb. 23, 2007	NIST releases pre-workshop results (ET results will not yet include scores that incorporate name variants)
Feb. 27, 2007	(noon deadline, EST) Site's detailed system description papers are due at NIST
Week of March 12th, 2007	NIST releases updated pre-workshop results for ET incorporating some of the name variants culled from system submissions
Mar. 28-29, 2007	Two day ACE07 evaluation workshop. (ET participants are invited)
Mar. 30, 2007	*One day Entity Translation workshop (ACE07 participants are invited)*
Apr. 20, 2007	Official public release of ACE07 and ACE ET07 results

[1] The official ACE07 registration form is located at the URL: http://www.nist.gov/speech/tests/ace/ace07/doc/

4.4 RULES

- Use of the ACE05 evaluation test set (source or reference) for any purpose whatsoever is prohibited.
- No changes to the system are allowed once the evaluation data are released. Adaptive systems may of course change themselves in response to the source data that they process.
- No human intervention is allowed prior to the submission of your test site's results to NIST.[2] This means that, in addition to disallowing modifications to your system, there must also be no modifications to, or human examination of, the test data.
- For each evaluation combination of task, language, and processing mode for which system output is submitted, all documents from all sources for that evaluation combination must be processed.
- Sites will receive the evaluation source data from NIST via email (see section 4.3 Schedule) and must return results to NIST before the end of the evaluation period.
- Every participating site must submit a detailed system description to NIST by Feb. 27, 2007, as defined in section 4.5.2.
- Every participating site must attend the evaluation workshop and present a system talk.

4.5 SUBMISSION OF SYSTEM OUTPUT TO NIST

To enable quick unpacking and scoring of several site submission files with minimum human intervention, participants must follow the outlined procedure for submitting results.

4.5.1 PACKAGING YOUR SYSTEM OUTPUT

STEP1: Create a top level directory for each of the ***languages*** attempted (Arabic and/or Chinese):

Example: $> mkdir arabic

STEP2: Create a subdirectory identifying the ***task*** (ET):

Example: $> mkdir arabic/et

STEP3: In each subdirectory make one directory for each system submitted (choose a name that identifies your site):

Example: $> mkdir arabic/et/NIST1_primary

Example: $> mkdir arabic/et/NIST2_contrastive

STEP4: Deposit all system output files in the appropriate system directory.

STEP5: Create a compressed tar file of your results and transfer them to NIST by FTP (ftp://jaguar.ncsl.nist.gov/incoming). After successful transmission send e-mail to ace_poc@nist.gov identifying the name of the file submitted. Alternatively you may send the compressed tar file directly to ace_poc@nist.gov .

4.5.2 SYSTEM DESCRIPTION

A valuable tool in discovering strengths and weakness of different algorithmic approaches is the use of system descriptions. System descriptions may also be used to help determine which sites are to give oral workshop presentations.

Each participant must prepare a ***detailed*** system description covering each system submitted. System descriptions are due at NIST no later than 02/27/07. It is important that all sites submit comprehensive descriptions on time so that NIST may plan the workshop agenda accordingly.

These system descriptions will be distributed to each participant before the evaluation workshop.

Each system description should include:

- The ET languages processed
- Identification of the primary system for each task
- A description of the system (algorithms, data, configuration) used to produce the system output
- How contrastive systems differ from the primary system
- A description of the resources required to process the test set, including CPU time and memory
- Applicable references

5 GUIDELINES FOR PUBLICATIONS

NIST Speech Group's HLT evaluations have been moving towards an open model which promotes interchange with the outside world. The rules governing the publication of the Entity Translation evaluation results are exactly the same as they are for ACE07.

5.1 NIST PUBLICATION OF RESULTS

At the conclusion of the evaluation cycle, NIST will create a report which documents the evaluation. The report will be posted on the NIST web space and will identify the participants and official ET value scores achieved for each track (language). Scores will be reported for the overall test set and for the different data sources. The NIST report will make it clear that this was a "pilot evaluation".

The report that NIST creates should not be construed, or represented as endorsements for any participant's system or commercial product, or as official findings on the part of NIST or the U.S. Government.

5.2 PARTICIPANT'S PUBLICATION OF RESULTS

Participants must refrain from publishing results and/or releasing statements of performance until the official ET07 results are posted by NIST on or around April 30th, 2007.

Participants may not compare its results with the results of other participants, such as stating rank ordering or score difference. Participants will be free to publish results for their own system, but, sites will not be allowed to name other participants, or cite

[2] It sometimes happens that a system bug is discovered during the course of processing the test data. In such a case, please consult with NIST via email (ace_poc@nist.gov) for advice. NIST will advise you on how to proceed. Repairs may be possible that allow a more accurate assessment of the underlying performance of a system. If this happens, modified results may be accepted, provided that an explanation of the modification is provided and provided that the original results are also submitted and documented.

another site's results without permission from the other site. Publications should point to the NIST report as a reference[3].

All publications must contain the following NIST disclaimer:

> ***NIST serves to coordinate the ET evaluations in order to support Entity Translation research and to help advance the state-of-the-art in entity translation technologies. ET evaluations are not viewed as a competition, as such, reported results by NIST are not to be construed, or represented, as endorsements of any participant's system, or as official findings on the part of NIST or the U.S. Government.***

[3] This restriction exists to ensure that readers concerned with a particular system's performance will see the entire set of participants and tasks attempted by all researchers.